Analisis Sentimen Masyarakat Terhadap Kinerja Komisi Pemilihan Umum (KPU) Menjelang Pemilu 2024 Menggunakan Metode *Support Vector Machine* Dan Seleksi Fitur *Information Gain*

Vita Rahmada[1], Anita Qoiriah[2]

[1,2]Program Studi S1 Teknik Informatika, Fakultas Teknik, Universitas Negeri Surabaya
[1]vita.20003@mhs.unesa.ac.id
[2]anitaqoiriah@unesa.ac.id

***Abstrak*— Kemajuan teknologi yang pesat telah mengubah cara individu berkomunikasi dan berinteraksi secara signifikan. Salah satu dampaknya adalah meningkatnya penggunaan media sosial, dengan media sosial X atau yang sebelumnya dikenal dengan Twitter yang sangat populer di Indonesia, menduduki peringkat keempat dunia dengan 25,25 juta pengguna pada Juli 2023 menurut laporan *We Are Social*. Menjelang pemilu 2024, kinerja Komisi Pemilihan Umum (KPU) menjadi sorotan publik, menjadikannya topik hangat di media sosial X. untuk memahami sentimen publik terhadap kinerja KPU, penelitian ini menggunakan metode *Support Vector Machine* (SVM), yang ditingkatkan dengan seleksi fitur Information Gain (IG), yang dikenal efektif dalam mengidentifikasi fitur relevan dalam dataset. Proses penelitian mencakup *preprocessing* data, pelabelan data, pembobotan TF-IDF, seleksi fitur dengan berbagai *threshold* (0.005, 0.007, 0.0005, 0.0007), lima skenario pembagian data untuk klasifikasi (50%:50%, 60%:40%, 70%:30%, 80%:20%, 90%:10%), dan evaluasi. Hasil menunjukkan bahwa SVM tanpa IG mencapai akurasi tertinggi sebesar 77,63%, sedangkan SVM dengan seleksi fitur IG mencapai akurasi sebesar 82,89% dengan *threshold* 0.0007. Kedua akurasi tersebut dicapai pada skenario pembagian data 90%:10%. Hal ini menunjukkan efektivitas seleksi fitur IG dalam meningkatkan akurasi klasifikasi sebesar 5,26%, serta meningkatkan nilai presisi, *recall*, dan *f-measure*. Dengan demikian, model SVM yang menggunakan seleksi fitur IG memberikan prediksi yang lebih akurat dan konsisten dalam menganalisis sentimen terhadap kinerja KPU menjelang pemilu 2024.**

Kata Kunci*— KPU, Pemilu, *Support Vector Machine*, Seleksi Fitur, *Information Gain.

I. PENDAHULUAN

Perkembangan teknologi merupakan fenomena yang tidak dapat dihindari pada saat ini. Perkembangan teknologi telah mengubah cara individu berkomunikasi dan berinteraksi satu sama lain. Salah satu dampak dari perkembangan teknologi adalah meningkatnya pengguna media sosial. Media sosial merupakan platform media yang berfokus pada partisipasi pengguna dan memfasilitasi kolaborasi di antara mereka [1]. Salah satu platform media sosial yang sangat populer saat ini adalah X atau yang dulunya dikenal dengan Twitter, hal ini dibuktikan dengan laporan *We Are Social* per Juli 2023 yang menunjukkan bahwa pengguna media sosial X di Indonesia berada di peringkat empat dunia dengan jumlah mencapai 25,25 juta pengguna [2]. Media sosial X memberikan kebebasan kepada pengguna untuk mengunggah postingan mengenai diri mereka sendiri maupun informasi yang dapat berguna untuk pengguna lainnya. Namun, banyak pengguna juga meninggalkan komentar negatif sebagai tanggapan atas suatu unggahan tertentu.

Media sosial X memiliki beberapa fitur menarik, salah satunya yaitu fitur hashtag. Melalui fitur tersebut, pengguna dapat menggunakan hashtag untuk mengkategorikan *tweet* mereka, sehingga dapat membuat pengguna lain dengan mudah menemukan topik yang serupa, *tweet* adalah sebutan khusus untuk postingan atau unggahan yang dibuat oleh pengguna. Saat ini, topik yang sering kali ramai diperbincangkan di media sosial X adalah topik mengenai kinerja komisi pemilihan umum (KPU) dalam pemilihan umum (Pemilu) 2024 mendatang. KPU merupakan lembaga penyelenggara pemilu yang diakui dan disahkan oleh UUD 1945, sehingga diklasifikasikan sebagai lembaga negara yang memiliki kedudukan penting secara konstitusional. KPU merupakan lembaga yang bersifat nasional, tetap, dan mandiri (independen). Mandiri atau independen dalam konteks ini yaitu penyelenggara pemilu harus terus netral dan tidak memihak [3]. Tahun 2024 mendatang akan menjadi tahun demokrasi terbesar bagi masyarakat Indonesia, oleh karena itu KPU yang merupakan institusi yang bertanggung jawab atas penyelenggaraan pemilu harus mempersiapkan diri dengan baik. Menjelang pemilu 2024, kinerja KPU menjadi sorotan di seluruh kalangan masyarakat, termasuk kalangan awam maupun *elite* politik sehingga sehingga KPU diharapkan terus meningkatkan kinerjanya demi memastikan proses pemilu berlangsung secara lancar, demokratis, dan adil [4].

Melalui situs media sosial X, masyarakat membagikan *tweet* yang berisi komentar dan opini mengenai kinerja KPU dalam mengelola pemilu di tahun 2024 mendatang. Seiring mendekati pelaksanaan pemilu, banyak isu-isu yang bermunculan terkait kinerja KPU sejak awal tahun 2023 seperti isu penundaan pemilu, keberadaan data pemilih yang mencurigakan, hingga isu ketiadaan debat calon presiden dan wakil presiden *pasca* pendaftaran bakal pasangan calon pemilu tahun 2024. Untuk mengetahui respon masyarakat terhadap topik kinerja KPU menjelang pemilu 2024 tersebut, perlu dilakukan analisis opini publik. Dengan melakukan analisis opini publik, dapat diketahui respon masyarakat

terhadap topik tersebut, serta bagaimana mengidentifikasi berita hoax, opini publik yang bersifat negatif, dan ujaran kebenjian terhadap pihak tertentu [5]. Salah satu metode analisis yang dapat diterapkan adalah analisis sentimen. Analisis sentimen merupakan bagian dari *Natural Language Processing* (NLP) yang berfokus pada mengkategorikan teks ke dalam tiga sentimen utama yaitu positif, negatif, dan netral. Proses analisis sentimen mencakup beberapa langkah penting, termasuk *preprocessing*, ekstraksi fitur, dan klasifikasi. Beberapa algoritma klasifikasi yang paling sering digunakan meliputi *Support Vector Machine* (SVM), *Naïve Bayes*, *Regresi Logistik*, *Random Forest*, dan *Decision Tree* [6].

Pada penelitian sebelumnya, Zain *et al*. melakukan eksplorasi metode terbaik menggunakan algoritma *Support Vector Machine* (SVM), *Naïve Bayes* dan *K-Nearest Neighbor* (KNN) untuk menganalisis sentimen twitter yang berkaitan dengan calon presiden 2024 [7]. Hasil penelitian tersebut menunjukkan bahwa algoritma SVM memberikan kinerja yang lebih baik dan tingkat konsistensi yang tinggi dalam presisi dan recall untuk setiap kategori sentimen. Beberapa penelitian serupa yang telah dilakukan terkait perbandingan algoritma dalam analisis sentimen antara lain adalah *Sentiment Analysis of Towards Electric Cars using Naive Bayes Classifier and Support Vector Machine Algorithm* [8], *Comparison Study of Sentiment Analysis of Tweets using Various Machine Learning Algorithms* [9], dan Analisis Sentimen Twitter Terhadap Steam Menggunakan Algoritma *Logistic Regression* dan *Support Vector Machine* [10]. Dari beberapa algoritma tersebut, algoritma yang paling sering mencapai hasil optimal adalah algoritma *Support Vector Machine* (SVM).

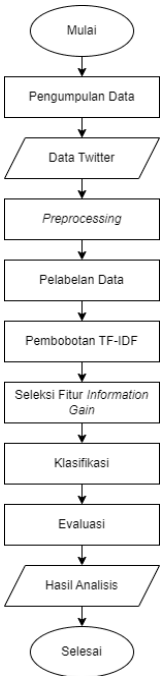
Namun, terdapat masalah dalam melakukan klasifikasi sentimen teks yaitu banyaknya jumlah atribut terdapat dalam dataset, yang dapat menyebabkan kinerja klasifikasi menurun. Untuk mengatasi masalah tersebut, seleksi fitur dapat diterapkan untuk mengidentifikasi dan menghapus atribut yang kurang relevan dalam dataset, dengan tujuan untuk meningkatkan akurasi klasifikasi. Sinanto Ate & Nuraminah melakukan penelitian dengan mengkomparasi algoritma seleksi fitur pada analisis sentimen review film [11]. Algoritma seleksi fitur yang dikomparasi antara lain *Information Gain* (IG), *Chi Square* (CS), *Forward Selection*, dan *Backward Elimination*, sedangkan algoritma klasifikasi yang digunakan yaitu *Support Vector Machine* (SVM), *Naïve Bayes* (NB), dan *Artificial Neural Network* (ANN). Hasil dari komparasi algoritma seleksi fitur tersebut menunjukkan bahwa IG merupakan algoritma seleksi fitur terbaik.

Penelitian dengan algoritma seleksi fitur serupa juga dilakukan oleh Madhumathi *et al*. yang melakukan analisis sentimen menggunakan pendekatan seleksi fitur berbasis Information Gain (IG) pada dataset yang berskala besar [12]. Hasil penelitian menunjukkan bahwa klasifikasi berbasis IG dapat meningkatkan akurasi analisis sentimen pada tingkat dokumen, kalimat, dan fitur.

Dengan demikian penelitian ini menerapkan metode *Support Vector Machine* sebagai klasifikasi dan *Information Gain* sebagai seleksi fitur untuk menganalisis sentimen masyarakat terhadap kinerja KPU menjelang pemilu 2024 melalui media sosial X.

II. METODE PENELITIAN

Penelitian ini memiliki alur atau tahapan mulai dari pengumpulan data sampai pembahasan hasil analisis seperti yang diuraikan pada Gbr. 1.



Gbr. 1 Alur Penelitian

Penjelasan dari alur penelitian tersebut adalah sebagai berikut.

A. Pengumpulan Data

Pengumpulan data dilakukan dengan cara *crawling tweet* dalam bahasa indonesia pada platform media sosial X dengan menggunakan *tweet-harvest*, yang merupakan sebuah alat yang digunakan untuk *crawling* data pada X. *Crawling tweet* pada penelitian ini menggunakan keyword "kpu". Data hasil *crawling* tersebut hanya diambil bagian komentar atau *tweet*nya saja yang kemudian disimpan dalam format csv dan diproses menggunakan bahasa pemrograman *Python*.

B. Preprocessing

Preprosessing adalah tahapan yang dilakukan dengan tujuan agar data mentah yang sudah diperoleh dapat diolah dan menghasilkan data yang terstruktur dan berkualitas. Berikut tahapan dalam melakukan *preprocessing*:

1) Cleaning

Proses *cleaning* pada penelitian ini melibatkan penghapusan komponen-komponen yang tidak diperlukan, seperti *Uniform Resource Locator* (URL), karakter HTML, username, emoji, angka, *hashtag*, dan RT (*Retweet*).

2) *Case Folding*

Proses ini bertujuan untuk mengubah semua huruf besar menjadi huruf kecil.

3) *Tokenization*

Proses ini dilakukan untuk memisahkan *tweet* menjadi potongan-potongan per kata.

4) *Normalization*

Proses ini bertujuan untuk menstandarisasi kata-kata dengan mengubah penulisan kata yang disingkat atau tidak baku menjadi bentuk yang lebih standar dan baku.

5) *Stopword Removal*

Selanjutnya, dilakukan proses penghapusan kata-kata yang tidak memiliki makna (*stopwords*).

6) *Stemming*

Terakhir, dilakukan proses stemming untuk mengubah kata berimbuhan menjadi kata dasar.

C. Pelabelan Data

Langkah berikutnya adalah memberikan label pada data tersebut agar mempermudah model dalam melakukan klasifikasi sentimen. Pada penelitian ini, pelabelan data dilakukan dengan menggunakan metode *Valence Aware Dictionary and Sentiment Reasoner* (VADER). VADER merupakan sebuah alat analisis yang memanfaatkan leksikon berbasis aturan dan pola gramatikal untuk menentukan sentimen teks. Namun, dalam penelitian ini, digunakan kamus *Indonesia Sentimen Lexicon* (InSet) dari repositori *GitHub*. Kamus InSet ini terdiri dari 3.609 kata positif dan 6.609 kata negatif dengan bobot sentimen mulai dari -5 hingga +5 (Fajri, 2017). Sebelum dilakukan pelabelan sentimen pada data *tweet*, leksikon *default* pada VADER dibersihkan terlebih dahulu.

TABEL 1

CONTOH DAFTAR KATA DALAM KAMUS INSET

Kata	Bobot	Kata	Bobot
Isak	-5	Muncul	1
Gila	-4	Hadir	2
Pening	-3	Untung	3
Asam	-2	Jujur	4
Isu	-1	Mantap	5

Selanjutnya, dilakukan pemuatan kamus InSet untuk sentimen positif dan negatif, lalu dilakukan pembaruan leksikon VADER dengan kamus InSet yang dimuat. Langkah berikutnya adalah menganalisis sentimen setiap *tweet* dengan menghitung skor sentimen positif, negatif, dan *compound score* menggunakan kamus InSet. *Compound score* tersebut kemudian digunakan untuk memberi label sentimen (positif atau negatif) kepada setiap *tweet* berdasarkan kamus InSet. Dalam pengelompokan, *tweet* dengan *compound score* lebih dari 0 akan diberi label positif, sedangkan *compound score* kurang dari atau sama dengan 0 akan diberi label negatif. Perhitungan *compound score* dilakukan menggunakan rumus pada persamaan 2.1.

$$\textit{Compound Score} = \frac{x}{\sqrt{x^2+\alpha}} \quad (2.1)$$

Pada persamaan (2.1), x merupakan jumlah nilai valensi kata penyusun dan α merupakan konstanta normalisasi dimana nilai *default*nya adalah 15.

D. Pembobotan TF-IDF

Pada penelitian ini, data yang berupa kumpulan kata tersebut akan dihitung bobotnya menggunakan persamaan TF-IDF. Kumpulan kata tersebut akan dijadikan sebagai fitur, dan semakin banyaknya dokumen yang diproses maka akan menghasilkan fitur yang lebih banyak. Pada tahap pembobotan TF-IDF, terdapat dua komponen utama yaitu TF (*Term Frequency*) dan IDF (*Inverse Document Frequency*). TF mengukur jumlah kemunculan kata dalam sebuah dokumen, dimana semakin banyak kata tersebut muncul, maka nilai TF akan semakin tinggi. Sementara itu, IDF mengacu pada jumlah nilai dokumen pada tiap kata, dimana kata yang jarang muncul akan memiliki nilai IDF yang lebih tinggi daripada kata yang sering muncul.

Persamaan untuk melakukan pembobotan TF-IDF adalah sebagai berikut:

$$TF_{ij} = \frac{f_{ij}}{max_k f_{kj}} \quad (2.2)$$

Pada persamaan (2.2), f_{ij} adalah jumlah munculnya term i dalam dokumen j, dan $max_k f_{kj}$ adalah total term pada dokumen j.

$$IDF_i = \log\left(\frac{N}{n_i}\right) \quad (2.3)$$

Pada persamaan (2.3), N adalah jumlah dokumen dan n_i adalah jumlah frekuensi dokumen untuk setiap term i.

$$TFIDF_{ij} = TF_{ij} \cdot IDF_i \quad (2.4)$$

Persamaan (2.4) adalah persamaan yang digunakan untuk menghitung nilai TF-IDF.

E. Seleksi Fitur Information Gain

Selanjutnya, dilakukan perhitungan *Information Gain* pada setiap fitur atau *term* dalam data *tweet*. Untuk mencari nilai *Information Gain*, langkah pertama adalah mencari nilai entropi. Berikut adalah persamaan untuk menghitung nilai entropi:

$$Entropy(S) = = -\sum_{i=1}^{c} p_i \cdot log_2 p_i \quad (2.5)$$

Pada persamaan (2.5), S merupakan sample, c merupakan jumlah kelas klasifikasi dan p_i merupakan proporsi sample di kelas i terhadap sample total.

$$Entropy_A(S) = \sum_{v \in Values(A)} \frac{|S_v|}{|S|} \cdot Entropy(S_v) \quad (2.6)$$

Pada persamaan (2.6), A adalah atribut, v adalah jumlah partisi dari atribut A, dan S_v adalah jumlah sample pada partisi ke v.

Setelah didapatkan nilai entropi, maka dapat dilakukan perhitungan nilai *Information Gain* dengan persamaan berikut:

$$Gain(S,A) = ntropy(S) - Entropy_A(S) \quad (2.7)$$

Dari perhitungan tersebut, nilai *Information Gain* yang lebih tinggi menunjukkan bahwa fitur tersebut lebih informatif dalam membedakan antara kelas-kelas dalam masalah

klasifikasi. Kemudian, dilakukan penentuan *threshold* untuk tahap seleksi fitur ini. Hasil fitur atau term yang terpilih sesuai dengan nilai *threshold* akan disimpan untuk digunakan ke tahap selanjutnya, sedangkan fitur yang tidak terpilih akan dihapus. Pada penelitian ini, digunakan *threshold* 0.005, 0.007, 0.0005, dan 0.0007 untuk menentukan fitur yang relevan dan signifikan dalam tahap seleksi fitur. *Threshold* pada *Information Gain* berfungsi untuk mencari data yang relevan dan menghapus data yang tidak terlalu penting [13].

F. Klasifikasi

Penelitian ini menggunakan metode *Support Vector Machine* (SVM) sebagai algoritma klasifikasi untuk menganalisis sentimen masyarakat terkait kinerja KPU menjelang pemilu 2024. Sebelum dilakukan klasifikasi, data yang telah melalui tahap *preprocessing* maupun seleksi fitur *Information Gain* akan dibagi menjadi data *training* dan data *testing*. Data *training* merupakan data yang akan digunakan untuk membangun dan melatih model klasifikasi. Data *testing* adalah data yang digunakan untuk melakukan pengujian terhadap hasil klasifikasi yang telah dibuat menggunakan data *training*. Pembagian data dilakukan dengan menggunakan lima skenario berbeda, yaitu 50:50%, 60%:40%, 70%:30%, 80%:20%, dan 90%:10% dengan tujuan untuk mencari kinerja terbaik dari metode atau model yang dihasilkan.

Setelah pembagian data, metode *Synthetic Minority Oversampling Technique* (SMOTE) diterapkan pada data *training* untuk menangani ketidakseimbangan data yang mungkin terjadi setelah pelabelan. SMOTE digunakan untuk menghasilkan data sintetik pada kelas minoritas sehingga jumlahnya sebanding dengan kelas mayoritas dan mencegah ketidakseimbangan yang dapat mempengaruhi kinerja model klasifikasi. Penerapan SMOTE setelah pembagian data pada data *training* bertujuan untuk memastikan bahwa data sintetik dihasilkan hanya berdasarkan data *training* sehingga membuat evaluasi model menjadi lebih realistis.

Pada penelitian ini, sebelum dilakukan klasifikasi menggunakan SVM, dilakukan *hyperparameter tuning* untuk membantu meningkatkan kinerja model dengan menemukan kombinasi parameter terbaik yang sesuai dengan data yang digunakan menggunakan metode *grid search*. Metode ini menguji semua kemungkinan kombinasi parameter yang telah ditentukan dalam sebuah *grid* untuk menentukan parameter yang optimal dan meningkatkan kinerja model [14].

G. Evaluasi

Tahap terakhir dari penelitian ini adalah evaluasi dengan menggunakan *confusion matrix* untuk menentukan nilai akurasi, presisi, *recall*, dan *f-measure*. Sebelum dilakukan perhitungan untuk mencari nilai-nilai tersebut, perlu dihitung terlebih dahulu nilai *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) pada data yang telah diklasifikasikan.

Berikut persamaan untuk mendapatkan nilai akurasi, presisi, *recall*, dan *f-measure*:

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.8)$$

$$\text{Presisi} = \frac{TP}{TP+FP} \quad (2.9)$$

$$\textit{Recall} = \frac{TP}{TP+FN} \quad (2.10)$$

$$\textit{F-measure} = \frac{2.\ presisi.\ recall}{presisi+recall} \quad (2.11)$$

Tujuan dari evaluasi ini adalah untuk mengetahui perbandingan akurasi model klasifikasi *Support Vector Machine* sebelum dan setelah penambahan seleksi fitur *Information Gain*. Melalui perbandingan akurasi tersebut, diharapkan dapat menentukan model terbaik untuk menganalisis sentimen masyarakat terhadap kinerja KPU menjelang pemilu 2024.

III. HASIL DAN PEMBAHASAN

A. Pengumpulan Data

Proses pengumpulan dataset yang digunakan untuk analisis sentimen ini dilakukan melalui proses *crawling* menggunakan *tweet*-harvest. Data yang dikumpulkan mencakup data dari tanggal 14 September 2023 hingga 13 Februari 2024 menggunakan *keyword* "kpu". Contoh data asli yang telah diperoleh dari proses *crawling* dan dapat dilihat pada Tabel 2.

TABEL 2

CONTOH DATA ASLI HASIL PROSES *CRAWLING*

No	***Tweet***
1	Laknat semua penyelenggara pemilu semua SDH di atur sama dinasti oligarki Bawaslu KPU simbol aja. 👎 👎 👎
2	Hati-hati diTIPU sm KPU lewat situng yg Seperti 2019 Sistem yg sudah tidak bisa dipercaya #WaktunyaPerubahan #AMIN2024
3	@KompasTV Rakyat yg memilih,,, survei yang menggiring opini ,,, KPU yg menentukan,,mudahan penyenggara pemilu nya bisa JURDIL,,,AMIN"

B. Preprocessing

Data *tweet* hasil *crawling* selanjutnya dilakukan proses *preprocessing* dengan melalui beberapa tahapan yaitu *cleaning*, *case folding*, tokenization, normalization, *stopword removal*, dan *stemming*. Data yang telah dilakukan *preprocessing* dapat dilihat pada Tabel 3.

TABEL 3

CONTOH DATA TWEET SETELAH *PREPROCESSING*

No	***Tweet* Setelah *Preprocessing***
1	['laknat', 'selenggara', 'pemilu', 'atur', 'dinasti', 'oligarki', 'bawaslu', 'kpu', 'simbol']
2	['hati', 'hati', 'tipu', 'kpu', 'situng', 'sistem', 'percaya']
3	['rakyat', 'pilih', 'survei', 'giring', 'opini', 'kpu', 'tentu', 'mudah', 'senggara', 'pemilu', 'nya', 'jujur adil', 'amin']

C. *Pelabelan Data*

Untuk mendapatkan nilai skor *compound* positif dan negatif pada Tabel 4, dilakukan perhitungan menggunakan rumus persamaan 2.1. Berikut hasil perhitungan skor *compound* positif dan negatif, serta total skor pada data *tweet* 1.

Skor compound positif = $\frac{3+4}{\sqrt{(3+4)^2+15}} = 0{,}88$

Skor compound negatif = $\frac{-5-4-3}{\sqrt{(-5-4-3)^2+15}} = -0{,}95$

Total skor = $0{,}88 + (-0{,}95) = -0{,}07$

TABEL 4

CONTOH PELABELAN DATA PADA DATA TWEET 1

Tweet Setelah Proses *Stemming*	Kata-kata Positif (Skor)	Kata-kata Negatif (Skor)	Skor Compound Positif	Skor Compound Negatif	Total Skor	Label
['laknat', 'selenggara', 'pemilu', 'atur', 'dinasti', 'oligarki', 'bawaslu', 'kpu', 'simbol']	selenggara (3), atur (4)	laknat (-5), atur (-4), dinasti (-3)	0,88	-0,95	-0,07	Negatif

Berdasarkan hasil total skor tersebut, karena total skor *compound* bernilai negatif, maka *tweet* tersebut termasuk dalam sentimen negatif. Hasil perhitungan pelabelan data menggunakan kamus InSet menunjukkan bahwa terdapat 1.013 data dengan sentimen negatif dan 502 data dengan sentimen positif.

D. *Pembobotan TF-IDF*

Pembobotan TF-IDF dimulai dengan menghitung nilai *Term Frequency* (TF). Nilai TF dihitung dengan menghitung jumlah kemunculan suatu *term* atau kata dalam suatu dokumen atau *tweet*, dibagi dengan total *term* dalam dokumen tersebut. Berikut contoh perhitungan TF pada data *tweet* 2 yaitu ['hati', 'hati', 'tipu', 'kpu', 'situng', 'sistem', 'percaya'] menggunakan rumus persamaan 2.2.

Langkah berikutnya adalah menghitung nilai *Inversed Document Frequency* (IDF) dengan menghitung jumlah dokumen yang mengandung setiap *term*, dibagi jumlah dokumen di mana suatu *term* muncul. Berikut contoh perhitungan IDF pada data *tweet* 2 menggunakan rumus persamaan 2.3.

Terakhir, nilai TF dan IDF dikalikan untuk mendapatkan nilai TF-IDF menggunakan rumus persamaan 2.4. Hasil pembobotan TF-IDF pada contoh data *tweet* hasil preprocessing dapat dilihat pada Tabel 5 dan Tabel 6.

TABEL 5

CONTOH PEMBOBOTAN TF-IDF

Term	T1	T2	T3	TF (T1)	TF (T2)	TF (T3)	IDF	TF-IDF (T1)	TF-IDF (T2)	TF-IDF (T3)
atur	1	0	0	0,125	0	0	0,477	0,060	0	0
bawaslu	1	0	0	0,125	0	0	0,477	0,060	0	0
dinasti	1	0	0	0,125	0	0	0,477	0,060	0	0
kpu	1	1	1	0,125	0,143	0,077	0	0	0	0
laknat	1	0	0	0,125	0	0	0,477	0,060	0	0
oligarki	1	0	0	0,125	0	0	0,477	0,060	0	0
pemilu	1	0	1	0,125	0	0,077	0,176	0,022	0	0,014
selenggara	1	0	0	0,125	0	0	0,477	0,060	0	0

TABEL 6

LANJUTAN CONTOH PEMBOBOTAN TF-IDF

Term	T1	T2	T3	TF (T1)	TF (T2)	TF (T3)	IDF	TF-IDF (T1)	TF-IDF (T2)	TF-IDF (T3)
hati	0	2	0	0	0,286	0	0,176	0	0,050	0
percaya	0	1	0	0	0,143	0	0,477	0	0,068	0
sistem	0	1	0	0	0,143	0	0,477	0	0,068	0
situng	0	1	0	0	0,143	0	0,477	0	0,068	0
tipu	0	1	0	0	0,143	0	0,477	0	0,068	0
adil	0	0	1	0	0	0,077	0,477	0	0	0,037
amin	0	0	1	0	0	0,077	0,477	0	0	0,037
giring	0	0	1	0	0	0,077	0,477	0	0	0,037
jujur	0	0	1	0	0	0,077	0,477	0	0	0,037
mudah	0	0	1	0	0	0,077	0,477	0	0	0,037
nya	0	0	1	0	0	0,077	0,477	0	0	0,037
opini	0	0	1	0	0	0,077	0,477	0	0	0,037
pilih	0	0	1	0	0	0,077	0,477	0	0	0,037
rakyat	0	0	1	0	0	0,077	0,477	0	0	0,037
survey	0	0	1	0	0	0,077	0,477	0	0	0,037
tentu	0	0	1	0	0	0,077	0,477	0	0	0,037

Tabel 5 dan Tabel 6 terdiri dari beberapa kolom, diantaranya adalah kolom “term”, yang berisi kata-kata atau fitur dari data *tweet* setelah melewati tahap preprocessing. Selain itu, terdapat kolom “T” yang menyimpan nilai jumlah kemunculan setiap term pada masing-masing data *tweet*, kolom “TF” yang menyimpan nilai term frequency, kolom “IDF” yang menyimpan nilai inverse document frequency, dan kolom “TF-IDF” yang memuat nilai TF-IDF dari setiap term atau fitur.

E. *Seleksi Fitur Information Gain*

Perhitungan *Information Gain* (IG) dilakukan menggunakan data hasil pelabelan. Data contoh tersebut terdiri dari 2 data *tweet* dengan kelas negatif dan 1 data *tweet* dengan kelas positif. Proses perhitungan ini dimulai dengan menghitung kemunculan fitur atau *term* pada masing-masing kelas. Tujuannya adalah untuk menentukan kelas yang sesuai untuk setiap *term* serta untuk menilai seberapa penting *term* tersebut. Hasil perhitungan kemunculan *term* pada tiap kelas ditunjukkan pada Tabel 7.

Selanjutnya adalah menghitung nilai IG pada tiap fitur atau *term*. Perhitungan nilai IG ini menggunakan contoh 3 data *tweet* yang terdiri dari 2 *tweet* dengan kelas negatif dan 1 *tweet* dengan kelas positif. Entropi dari dataset tersebut dihitung menggunakan rumus persamaan 2.5 sebagai berikut:

Entropy(Positif) = $-\left(\frac{1}{3}log_2\frac{1}{3}\right) = -\left(\frac{1}{3}x - 0{,}477\right) = 0{,}159$

Entropy(Negatif = $-\left(\frac{2}{3}log_2\frac{2}{3}\right) = -\left(\frac{2}{3}x - 0{,}176\right) = 0{,}117$

Entropy(S) = $0{,}159 + 0{,}117 = 0{,}276$

TABEL 7

CONTOH PERHITUNGAN KEMUNCULAN TERM

Term	Positif	Negatif	*Term*	Positif	Negatif
atur	0	1	tipu	0	1
bawaslu	0	1	adil	1	0
dinasti	0	1	amin	1	0
kpu	1	2	giring	1	0
laknat	0	1	jujur	1	0
oligarki	0	1	mudah	1	0
pemilu	1	1	nya	1	0
selenggara	0	1	opini	1	0
hati	0	1	pilih	1	0
percaya	0	1	rakyat	1	0
sistem	0	1	survey	1	0
situng	0	1	tentu	1	0

Selanjutnya, dilakukan perhitungan nilai entropi dari kata "pemilu", dimana kata tersebut muncul dalam 2 data *tweet* yaitu 1 *tweet* dengan kelas positif dan 1 *tweet* dengan kelas negatif, serta kata tersebut tidak muncul dalam 1 *tweet* dengan kelas negatif menggunakan rumus persamaan 2.6 sebagai berikut:

Entropy(Pemilu Muncul) = $-\left(\frac{1}{2}log_2\frac{1}{2}\right) - \left(\frac{1}{2}log_2\frac{1}{2}\right) = -\left(\frac{1}{2}x - 0{,}301\right) - \left(\frac{1}{2}x - 0{,}301\right) = 0{,}151 + 0{,}151 = 0{,}302$

Entropy(Pemilu Tidak Muncul) = $-(1log_2 1) = 0$

Entropy(S) = $\left(\frac{2}{3}x\ 0{,}302\right) + \left(\frac{1}{3}x\ 0\right) = 0{,}201$

Langkah terakhir yaitu menghitung nilai IG dari kata pemilu dengan menggunakan rumus persamaan 2.7 sebagai berikut:

Gain(S,A) = $0{,}276 - 0{,}201 = 0{,}075$

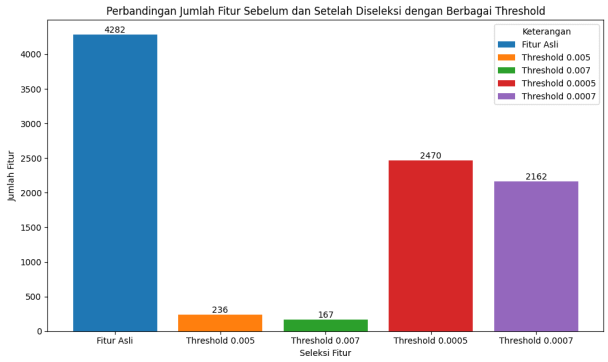
Perhitungan nilai IG dilakukan pada setiap fitur atau *term* dalam dataset. Hasil perhitungan pada setiap fitur kemudian diurutkan berdasarkan nilai IG teratas. Hasil perhitungan 10 nilai IG teratas dapat dilihat pada Tabel 8.

TABEL 8

10 NILAI IG TERATAS

Term	Information Gain
kpu	0,627
presiden	0,097
kah	0,094
bawaslu	0,082
calon	0,080
debat	0,075
pemilu	0,075
iya	0,059
daftar	0,056
nya	0,055

Dari Tabel 8, terlihat bahwa term atau fitur yang memiliki nilai IG tertinggi adalah "kpu" dengan nilai IG sebesar 0.627. Selanjutnya, kolom nilai *Information Gain* pada Tabel 8 akan difilter dengan menggunakan *threshold* 0.005, 0.007, 0.0005, dan 0.0007. Fitur-fitur dengan nilai IG kurang dari atau sama dengan *threshold* tersebut akan dihapus, sedangkan fitur-fitur dengan nilai IG lebih besar dari *threshold* akan dipertahankan untuk tahap berikutnya. Jumlah fitur sebelum dilakukan seleksi fitur IG adalah 4282. Sedangkan jumlah fitur setelah dilakukan seleksi fitur dengan *threshold* 0.005, 0.007, 0.0005, dan 0.0007 masing-masing adalah 236, 167, 2470, dan 2162 fitur. Perbandingan jumlah fitur sebelum dan setelah dilakukan seleksi fitur IG berdasarkan *threshold* tersebut ditampilkan pada Gbr. 2.



Gbr. 2 Perbandingan Jumlah Fitur Sebelum dan Setelah Seleksi Fitur

F. *Klasifikasi*

1) Pembagian Data

Proses klasifikasi dimulai dengan membagi data menjadi data *training* dan data *testing*. Tujuannya adalah untuk memastikan bahwa model yang dibangun dapat

dievaluasi dengan baik. Rincian pembagian data dapat dilihat pada Tabel 9.

TABEL 9

PEMBAGIAN DATA *TRAINING* DAN DATA *TESTING*

Jumlah Data *Tweet*	Pembagian Data *Training* dan Data *Testing*									
	Train	***Test***	***Train***	***Test***	***Train***	***Test***	***Train***	***Test***	***Train***	***Test***
	50%	**50%**	**60%**	**40%**	**70%**	**30%**	**80%**	**20%**	**90%**	**10%**
1515	757	758	909	606	1060	455	1212	303	1363	152

Tabel 9 memberikan detail pembagian data *training* dan data *testing* dalam lima skenario yang berbeda, dengan presentase pembagian yang bervariasi. Dari total 1515 data *tweet* tersebut, terdapat ketidakseimbangan kelas dengan 1013 data negatif dan 502 data positif. Ketidakseimbangan ini dapat mempengaruhi kinerja model klasifikasi, terutama dalam mengklasifikasikan data dari kelas minoritas, yaitu kelas positif. Untuk mengatasi masalah ketidakseimbangan kelas tersebut, digunakan metode *Synthetic Minority Over-sampling Technique* (SMOTE).

Implementasi SMOTE dilakukan menggunakan fungsi *fit_resample()* dari kelas SMOTE yang disediakan oleh *library imblearn*, yang merupakan *library* pada *Python* yang membantu menangani masalah ketidakseimbangan kelas dalam dataset. Hasil dari proses SMOTE adalah dataset yang disesuaikan sehingga kelas minoritas memiliki jumlah yang sama dengan kelas mayoritas. Perbandingan jumlah data negatif dan positif pada data *training* sebelum dan setelah dilakukan proses SMOTE pada masing-masing skenario pembagian data dapat dilihat pada Tabel 10 berikut.

TABEL 10

PERBANDINGAN JUMLAH DATA *TRAINING* SEBELUM DAN SETELAH SMOTE

Perbandingan Jumlah Data *Training*	Skenario 50%:50%		Skenario 60%:40%		Skenario 70%:30%		Skenario 80%:20%		Skenario 90%:10%	
	Pos	**Neg**	**Pos**	**Neg**	**Pos**	**Neg**	**Pos**	**Neg**	**Pos**	**Neg**
Sebelum SMOTE	253	504	293	616	343	717	396	816	447	916
Setelah SMOTE	504	504	616	616	717	717	816	816	916	916

2) Hyperparameter Tuning

Langkah selanjutnya adalah melakukan klasifikasi menggunakan algoritma *Support Vector Machine* (SVM). Agar model SVM dapat bekerja secara optimal, dilakukan *hyperparameter tuning* untuk menemukan kombinasi parameter terbaik yang sesuai dengan data yang digunakan. Metode *grid search* diterapkan untuk mencari kombinasi hyperparameter terbaik dengan menguji semua kemungkinan kombinasi dari set *hyperparameter* yang telah ditentukan dalam sebuah *grid*. Metode *grid search* ini diimplementasikan menggunakan fungsi *GridSearchCV()* dari *library scikit-learn*, yang memungkinkan pencarian menyeluruh terhadap kombinasi parameter yang berbeda untuk menemukan kombinasi parameter terbaik.

```
Akurasi :  0.9034923142753695
Parameter Terbaik : {'C': 10, 'gamma': 0.1, 'kernel': 'rbf'}
```

Gbr. 3. Hasil Hyperparameter Tuning

Gbr. 3 menunjukkan hasil dari *hyperparameter tuning* menggunakan metode *grid search*. Kombinasi *hyperparameter* terbaik yang diperoleh dalam penelitian ini adalah nilai C sebesar 10, nilai *gamma* sebesar 0.1, dan *kernel* RBF.

3) Pengujian Model tanpa Seleksi Fitur *Information Gain*

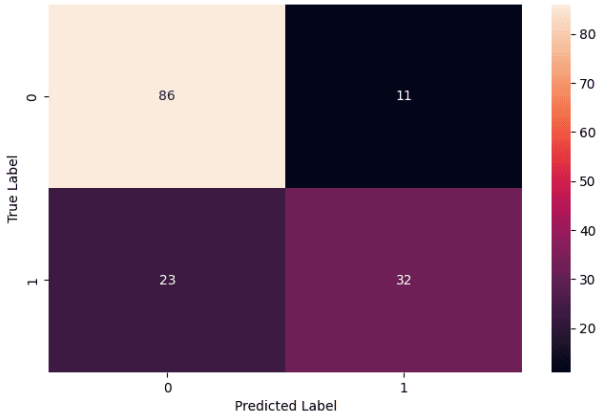
Pengujian pertama dilakukan menggunakan model klasifikasi SVM tanpa seleksi fitur *Information Gain*. Tujuannya adalah untuk mengevaluasi kinerja dari model SVM itu sendiri, yang kemudian akan digunakan sebagai pembanding untuk pengujian selanjutnya.

TABEL 11

HASIL PENGUJIAN SVM TANPA SELEKSI FITUR IG

Data *Training* : Data *Testing* (%)	**Akurasi**	**Presisi**	***Recall***	***F-measure***
50 : 50	74,54%	65,56%	47,39%	55,01%
60 : 40	75,08%	71,01%	46,89%	56,48%
70 : 30	76,26%	71,43%	53,46%	61,15%
80 : 20	77,23%	72,84%	55,66%	63,10%
90 : 10	**77,63%**	**74,42%**	**58,18%**	**65,31%**

Tabel 11 menampilkan hasil pengujian pertama, menunjukkan bahwa nilai akurasi tertinggi diperoleh ketika presentase data *training* adalah 90% dan data *testing* adalah 10%, dengan nilai akurasi mencapai 77,63%. Selain itu, terdapat nilai presisi sebesar 74,42%, *recall* sebesar 58,18%, dan *f-measure* sebesar 65,31%. Lebih lanjut, detail hasil *confusion matrix* dari penggunaan metode SVM tanpa seleksi fitur *Information Gain* pada rasio 90%:10% dapat dilihat pada Gbr. 4.



Gbr. 4. Hasil *Confusion Matrix* Tanpa Seleksi Fitur

Akurasi = $\frac{TP+TN}{TP+TN+FP+FN} x100\% = \frac{32+86}{32+86+11+23} x100\% = 77,63\%$

Presisi Positif = $\frac{TP}{TP+FP} x100\% = \frac{32}{32+11} x100\% = 74,42\%$

Presisi Negatif = $\frac{TN}{TN+FN} x100\% = \frac{86}{86+23} x100\% = 78,90\%$

Recall Positif = $\frac{TP}{TP+FN} x100\% = \frac{32}{32+23} x100\% = 58,18\%$

Recall Negatif = $\frac{TN}{TN+FP} x100\% = \frac{86}{86+11} x100\% = 88,66\%$

F-measure Positif = $2x\frac{presisi\ positif\ x\ recall\ positif}{presisi\ positif + recall\ positif} = 2x\frac{0,7442\ x\ 0,5818}{0,7442 + 0,5818} = 65,31\%$

F-measure Negatif = $2x\frac{presisi\ negatif\ x\ recall\ negatif}{presisi\ negatif + recall\ negatif} = 2x\frac{0,7890\ x\ 0,8866}{0,7890 + 0,8866} = 83,50\%$

4) Pengujian Model dengan Seleksi Fitur *Information Gain*

Pengujian berikutnya adalah pengujian menggunakan model klasifikasi SVM dengan data yang telah melalui seleksi fitur *Information Gain* pada tahap sebelumnya. Tujuan pengujian ini adalah untuk mengevaluasi kinerja dari model klasifikasi, yang dimana hasil dari pengujian tersebut akan dibandingkan dengan pengujian sebelumnya.

TABEL 12

HASIL PENGUJIAN SVM DENGAN SELEKSI FITUR IG

Data *Training* : Data *Testing* (%)	***Threshold***	**Jumlah Fitur**	**Akurasi**	**Presisi**	***Recall***	***F-measure***
50 : 50	0.005	236	73,35%	61,81%	49,40%	54,91%
	0.007	167	73,75%	62,89%	49,00%	55,08%
	0.0005	2470	77,31%	72,78%	49,40%	58,85%
	0.0007	2162	79,68%	75,96%	55,82%	64,35%
60 : 40	0.005	236	72,61%	62,15%	52,63%	56,99%
	0.007	167	72,28%	62,28%	49.76%	55,32%
	0.0005	2470	77,56%	77,04%	49,76%	60,47%
	0.0007	2162	79,21%	79,86%	53,11%	63,79%
70 : 30	0.005	236	74,07%	64,54%	57,23%	60,67%
	0.007	167	73,41%	63,97%	54,72%	58,98%
	0.0005	2470	79,78%	79,65%	56,60%	66,18%
	0.0007	2162	81,32%	79,84%	62,26%	69,96%
80 : 20	0.005	236	75,58%	67,02%	59,43%	63,00%
	0.007	167	73,60%	62,50%	61,32%	61,90%
	0.0005	2470	78,88%	76,25%	57,55%	65,59%
	0.0007	2162	80,86%	78,57%	62,26%	69,47%
90 : 10	0.005	236	78,29%	73,91%	61,82%	67,33%
	0.007	167	78,29%	72,92%	63,64%	67,96%
	0.0005	2470	80,92%	80,95%	61,82%	70,10%
	0.0007	**2162**	**82,89%**	**82,22%**	**67,27%**	**74,00%**

Tabel 12 menampilkan hasil pengujian kedua, menunjukkan bahwa nilai akurasi tertinggi diperoleh ketika presentase data *training* adalah 90% dan data *testing* adalah 10%, dengan *threshold* sebesar 0.0007 dan jumlah fitur sebanyak 2162 fitur. Akurasi yang diperoleh mencapai 82,89%. Selain itu, terdapat nilai presisi sebesar 82,22%, *recall* sebesar 67,27%, dan *f-measure* sebesar 74,00%. Lebih lanjut, detail hasil *confusion matrix* dari penggunaan metode SVM dengan seleksi fitur *Information Gain* pada rasio 90%:10% dapat dilihat pada Gbr. 5.

Akurasi = $\frac{TP+TN}{TP+TN+FP+FN} x100\% = \frac{37+89}{37+89+8+18} x100\% = 82,89\%$

Presisi Positif = $\frac{TP}{TP+FP} x100\% = \frac{37}{37+8} x100\% = 82,22\%$

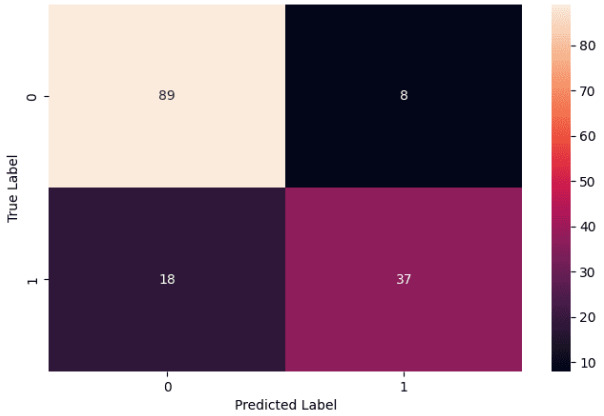
Presisi Negatif = $\frac{TN}{TN+FN} x100\% = \frac{89}{89+18} x100\% = 83,18\%$

Recall Positif = $\frac{TP}{TP+FN} x100\% = \frac{37}{37+18} x100\% = 67,27\%$

Recall Negatif = $\frac{TN}{TN+FP} x100\% = \frac{89}{89+8} x100\% = 91,75\%$

F-measure Positif = $2x\frac{presisi\ positif\ x\ recall\ positif}{presisi\ positif + recall\ positif} = 2x\frac{0,8222\ x\ 0,6727}{0,8222 + 0,6727} x100\% = 74,00\%$

F-measure Negatif = $2x\frac{presisi\ negatif\ x\ recall\ negatif}{presisi\ negatif + recall\ negatif} = 2x\frac{0,8318\ x\ 0,9175}{0,8318 + 0,9175} x100\% = 87,26\%$



Gbr. 5. Hasil Confusion Matrix dengan Seleksi Fitur

G. *Visualisasi Word Cloud*

Visualisasi *word cloud* bertujuan untuk memberikan gambaran kata-kata yang paling sering muncul dari kedua kategori sentimen yaitu sentimen positif dan negatif, sehingga memudahkan dalam memahami pola dan karakteristik dari data yang dianalisis. Kata-kata yang sering muncul dari sentimen positif dan negatif diidentifikasi dan divisualisasikan untuk memberikan gambaran yang lebih jelas tentang sentimen masyarakat terhadap KPU menjelang pemilu 2024. Gbr. 6 dan Gbr. 7 merupakan hasil visualisasi word cloud untuk masing-masing sentimen.

Hasil visualisasi *word cloud* menunjukkan bahwa kata-kata yang sering muncul dari kedua sentimen tersebut adalah "kpu", "pemilu", "bawaslu", "pilih", dan "debat". kata-kata ini menggambarkan topik utama yang menjadi perhatian masyarakat terkait kinerja KPU menjelang pemilu 2024. Kata "kpu" dan "pemilu" secara langsung merujuk pada lembaga dan perihal utama yang menjadi fokus analisis, sementara "bawaslu" menunjukkan adanya perhatian terhadap lembaga pengawas pemilu. Kata "pilih" dan "debat" menunjukkan kegiatan utama yang terjadi dalam proses pemilu, yaitu pemilihan dan debat kandidat.

Gbr. 6. *Word Cloud* Sentimen Positif

Selanjutnya, beberapa kata yang muncul dari kelas positif adalah "dukung", "menang", "kampanye", "atur", dan lain sebagainya. Kata "dukung" menunjukkan adanya sentimen positif dari masyarakat yang mendukung kinerja KPU, sedangkan kata "menang" dan "kampanye" menggambarkan harapan positif dan dukungan terhadap proses kampanye dan hasil pemilu yang adil. Kata "atur" menunjukkan apresiasi terhadap upaya KPU dalam mengatur dan menyelenggarakan pemilu secara baik.

Gbr. 7. *Word Cloud* Sentimen Negatif

Sebaliknya, beberapa kata yang muncul dari kelas negatif adalah "curang", "salah", "takut", "kalah", dan lain sebagainya. Kata "curang" dan "salah" menggambarkan ketidakpuasan masyarakat terhadap adanya dugaan kecurangan atau kesalahan dalam proses pemilu. Kata "takut" dan "kalah" menunjukkan adanya kekhawatiran dan ketakutan akan hasil pemilu yang tidak adil atau adanya pihak yang merasa dirugikan. Melalui kata-kata yang sering muncul tersebut, KPU dapat mempertimbangkan perbaikan kinerja dalam pelaksanaan pemilu berikutnya.

H. *Pembahasan*

Hasil analisis dari penelitian ini mencakup proses pengumpulan data melalui proses *crawling* pada platform sosial media X atau Twitter dalam rentang waktu 14 September hingga 13 Februari 2024. Total data yang berhasil dikumpulkan sebanyak 1515 data *tweet*. Selanjutnya, analisis dilakukan menggunakan bahasa pemrograman *Python*. Data yang telah terkumpul melalui proses *crawling* tersebut dilakukan serangkaian tahapan *preprocessing* yang meliputi *cleaning*, *case folding*, *tokenization*, *normalization*, *stopword removal*, dan *stemming*. Setelah *preprocessing*, data diberi label menggunakan kamus *Indonesia Sentimen Lexicon* (InSet). Hasilnya, terdapat 1013 data dengan sentimen negatif dan 502 data dengan sentimen positif. Beberapa kata yang sering muncul dalam kedua sentimen tersebut antara lain kpu, pemilu, bawaslu, cawapres, calon presiden, dan debat. Selain kata-kata tersebut, beberapa kata yang muncul dalam sentimen positif adalah dukung, percaya, menang, pilih, jujur, adil, dan lain sebagainya. Sedangkan beberapa kata yang muncul dalam sentimen negatif adalah curang, kalah, lawan, langgar, salah, isu bohong, dan lain sebagainya.

Setelah tahap *preprocessing* dan pelabelan data, dilakukan perhitungan TF-IDF dan dilanjutkan dengan seleksi fitur menggunakan *Information Gain* (IG). Dengan menerapkan *threshold* sebesar 0.005, 0.007, 0.0005, dan 0.0007, jumlah fitur berhasil direduksi dari 4282 menjadi masing-masing sebesar 236, 167, 2470, dan 2162 fitur. Selanjutnya, dilakukan pembagian data dan penerapan metode SMOTE untuk menangani ketidakseimbangan data, yang diterapkan khusus pada data *training*. Setelah itu, dilakukan hyperparameter tuning untuk mencari kombinasi parameter terbaik dalam

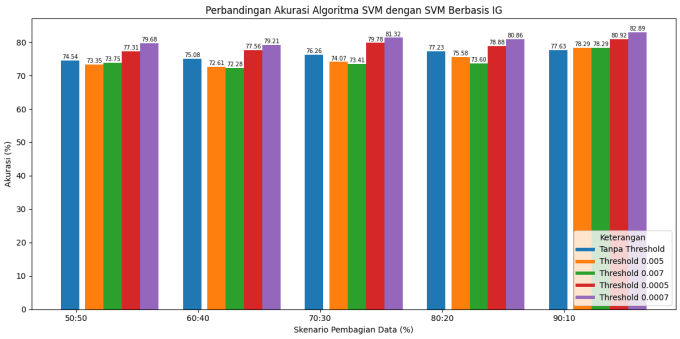
melakukan klasifikasi. Dalam proses ini, didapatkan kombinasi parameter optimal dengan nilai C sebesar 10, nilai *gamma* sebesar 0.1, dan *kernel* RBF. Hasil dari pengujian menggunakan lima skenario pembagian dataset, yaitu 70%:30%, 80%:20%, dan 90%:10%, serta menggunakan kombinasi parameter terbaik tersebut dapat dilihat pada Tabel berikut.

TABEL 13

HASIL PERBANDINGAN ALGORITMA SVM DENGAN SVM BERBASIS IG

Sk.	Pengujian 1 (SVM)				*Thres-hold*	Jml F	Pengujian 2 (SVM + IG)				*Delta*	Hasil
	Akur.	Pres.	*Rec.*	*F-Meas*			Akur.	Pres.	*Rec.*	*F-Meas*		
50 : 50	74,54	65,56	47,39	55,01	0.005	236	73,35	61,81	49,40	54,91	-1,19	Turun
					0.007	167	73,75	62,89	49,00	55,08	-0,79	Turun
					0.0005	2470	77,31	72,78	49,40	58,85	2,77	Naik
					0.0007	2612	79,68	75,96	55,82	64,35	5,14	Naik
60 : 40	75,08	71,01	46,89	56,48	0.005	236	72,61	62,15	52,63	56,99	-2,47	Turun
					0.007	167	72,28	62,28	49.76	55,32	-2,80	Turun
					0.0005	2470	77,56	77,04	49,76	60,47	2,48	Naik
					0.0007	2612	79,21	79,86	53,11	63,79	4,13	Naik
70 : 30	76,26	71,43	53,46	61,15	0.005	236	74,07	64,54	57,23	60,67	-2,19	Turun
					0.007	167	73,41	63,97	54,72	58,98	-2,85	Turun
					0.0005	2470	79,78	79,65	56,60	66,18	3,52	Naik
					0.0007	2612	81,32	79,84	62,26	69,96	5,06	Naik
80 : 20	77,23	72,84	55,66	63,10	0.005	236	75,58	67,02	59,43	63,00	-1,65	Turun
					0.007	167	73,60	62,50	61,32	61,90	-3,63	Turun
					0.0005	2470	78,88	76,25	57,55	65,59	1,65	Naik
					0.0007	2162	80,86	78,57	62,26	69,47	3,63	Naik
90 : 10	**77,63**	**74,42**	**58,18**	**65,31**	0.005	236	78,29	73,91	61,82	67,33	0,66	Naik
					0.007	167	78,29	72,92	63,64	67,96	0,66	Naik
					0.0005	2470	80,92	80,95	61,82	70,10	3,29	Naik
					0.0007	**2612**	**82,89**	**82,22**	**67,27**	**74,00**	**5,26**	Naik

Tabel 13 menampilkan hasil perbandingan antara algoritma SVM dengan SVM berbasis IG dalam berbagai skenario pengujian. Kolom “Sk.” merupakan singkatan dari Skenario yang menunjukkan rasio pembagian data *training* dan data *testing*. “Ak.” adalah singkatan dari Akurasi, “Pres.” adalah singkatan dari Presisi, “Rec.” adalah singkatan dari *Recall*, dan “F-Meas” adalah singkatan dari *F-measure*. Selanjutnya, kolom “Jml F” merupakan jumlah fitur dari masing-masing *threshold* yang ada di kolom *threshold*. Kolom “Delta” menunjukkan selisih persentase antara akurasi SVM berbasis IG dengan SVM. Kolom “Hasil” menunjukkan keterangan dari kolom delta, apakah persentase akurasinya naik atau turun. Selain itu, perbandingan akurasi dari kedua pengujian tersebut juga ditampilkan pada Gbr 8.



Gbr. 8. Perbandingan Akurasi Algoritma SVM dengan SVM Berbasis IG

Berdasarkan hasil perbandingan, nilai akurasi tertinggi pada pengujian menggunakan model SVM adalah sebesar 77,63% dengan nilai presisi sebesar 74,42%, *recall* sebesar 58,18%, dan *f-measure* sebesar 65,31%. Hasil tersebut diperoleh ketika skenario data *training* dan data *testing* adalah 90%:10%. Sementara itu, pengujian menggunakan model SVM dengan seleksi fitur IG menunjukkan nilai akurasi tertinggi sebesar 82,89%, dengan nilai presisi sebesar presisi sebesar 82,22%, *recall* sebesar 67,27%, dan *f-measure* sebesar 74,00%. Akurasi tertinggi pada pengujian tersebut didapatkan ketika *threshold* bernilai 0.0007 dengan skenario data *training* dan data *testing* adalah 90%:10%. Selisih peningkatan akurasi tertinggi tersebut adalah sebesar 5,26%. Hal ini menunjukkan bahwa seleksi fitur IG memberikan pengaruh yang signifikan terhadap peningkatan kinerja model.

Peningkatan akurasi terus terjadi ketika nilai *threshold* sebesar 0.0005 dan 0.0007. Hal ini menunjukkan bahwa menggunakan *threshold* yang lebih kecil dapat menyeleksi fitur-fitur yang lebih relevan dan signifikan, sehingga meningkatkan kinerja model. Namun, pada *threshold* 0.005 dan 0.007, akurasi menurun, yang berarti bahwa penggunaan *threshold* yang terlalu tinggi mungkin menyeleksi terlalu banyak fitur penting, sehingga mengurangi kinerja model. Selain itu, hasil kedua pengujian juga menunjukkan bahwa nilai presisi, *recall*, dan *f-measure* meningkat seiring dengan meningkatnya nilai akurasi, menandakan bahwa model menjadi lebih konsisten dalam prediksi sentimen.

IV. KESIMPULAN

Implementasi metode klasifikasi SVM dengan seleksi fitur Information Gain untuk analisis sentimen masyarakat terhadap kinerja KPU menjelang pemilu 2024 dilakukan melalui beberapa tahapan. Tahapan tersebut meliputi *preprocessing* data, pelabelan data, pembobotan TF-IDF, seleksi fitur Information Gain dengan penerapan *threshold* yang bervariasi (0.005, 0.007, 0.0005, dan 0.0007), pembagian data dengan lima skenario pembagian dataset (50%:50%, 60%:40%, 70%:30%, 80%:20%, dan 90%:10%), penerapan SMOTE, *hyperparameter tuning*, klasifikasi menggunakan SVM, dan evaluasi model.

Hasil klasifikasi menggunakan metode SVM dengan parameter optimal (nilai C sebesar 10, nilai *gamma* sebesar 0.1, dan *kernel* RBF) pada data yang terdiri dari 1013 data

dengan sentimen negatif dan 502 data dengan sentimen positif menunjukkan akurasi tertinggi 77,63% dengan presentase data *training* adalah 90% dan data *testing* adalah 10%. Nilai presisi prediksi positif mencapai 74,42%, presisi prediksi negatif 78,90%, *recall* data positif 58,18%, *recall* data negatif 88,66%, *f-measure* positif 65,31%, dan *f-measure* negatif 83,50%. Sedangkan hasil klasifikasi menggunakan metode SVM berbasis seleksi fitur *Information Gain* menunjukkan akurasi tertinggi sebesar 82,89% dengan *threshold* 0.0007 serta presentase data *training* adalah 90% dan data *testing* adalah 10%. Nilai presisi prediksi positif mencapai 82,22%, presisi prediksi negatif 83,18%, *recall* data positif 67,27%, *recall* data negatif 91,75%, *f-measure* positif 74,00%, dan *f-measure* negatif 87,26%. Berdasarkan hasil tersebut, penggunaan seleksi fitur *Information Gain* pada SVM terbukti efektif dalam meningkatkan kinerja klasifikasi. Dengan *threshold* 0.0007, model SVM yang menggunakan seleksi fitur Information Gain menunjukkan peningkatan akurasi sebesar 5,26% dibandingkan model SVM tanpa seleksi fitur. Peningkatan ini juga diiringi dengan peningkatan nilai presisi, *recall*, dan *f-measure*, yang menunjukkan bahwa model SVM dengan seleksi fitur *Information Gain* mampu menghasilkan prediksi yang lebih akurat dan konsisten dalam menganalisis sentimen masyarakat terhadap kinerja KPU menjelang pemilu 2024.

V. SARAN

Berdasarkan hasil penelitian ini, terdapat beberapa saran untuk pengembangan penelitian selanjutnya yaitu:

1. Dataset yang dikumpulkan dapat diperluas sehingga tidak hanya menggunakan data dari platform sosial media X saja.
2. Pelabelan data pada penelitian ini menggunakan kamus InSet yang masih terbatas dan belum lengkap. Untuk penelitian selanjutnya disarankan menggunakan kamus selain InSet atau mengembangkan kamus untuk meningkatkan akurasi pelabelan data.
3. Penelitian selanjutnya dapat menggunakan metode seleksi fitur lainnya untuk mengetahui kinerja dari metode seleksi fitur lain tersebut dalam menganalisis sentimen.

UCAPAN TERIMA KASIH

Segala puji dan syukur penulis panjatkan kepada Allah SWT yang telah memberikan kesehatan, ilmu, serta kemudahan dalam proses penyusunan artikel ilmiah ini. Penulis sangat berterima kasih kepada kedua orang tua dan keluarga yang selalu memberikan dukungan dan doa tanpa henti. Terima kasih juga penulis sampaikan kepada sahabat-sahabat yang telah menemani, memberikan masukan, serta membantu selama penelitian ini berlangsung. Tak lupa, penulis juga mengapresiasi diri sendiri yang telah kuat, bertahan, dan berjuang hingga artikel ilmiah ini dapat terselesaikan dengan baik.

REFERENSI

[1] Nurlaely, R., Simatupang, D. S., Kamdan, & Kharisma, I. L. (2023). Analisis Sentimen Twitter Terhadap Cyberbullying Menggunakan Metode Support Vector Machine (SVM). *Jurnal Computer Science and Information Technology (CoSciTech)*, *4*(2), 376–384.

[2] Annur, C. M. (2023). *Jumlah Pengguna Twitter Indonesia Duduki Peringkat ke-4 Dunia per Juli 2023*. Databoks. https://databoks.katadata.co.id/datapublish/2023/11/01/jumlah-pengguna-twitter-indonesia-duduki-peringkat-ke-4-dunia-per-juli-2023

[3] Syawawi, R. (2020). Konstitusionalitas Kemandirian Lembaga Penyelenggara Pemilihan Umum (Analisis Putusan Mahkamah Konstitusi Nomor 81/PUU- IX/2011). *Jurnal Ilmiah Penegakan Hukum*, *7*(2), 98–116.

[4] Amirullah, F., Alam, S., & Sulistyo S., M. I. (2023). Analisis Sentimen Terhadap Kinerja KPU Menjelang Pemilu 2024 Berdasarkan Opini Twitter Menggunakan Naïve Bayes. *Jurnal Ilmiah Teknik Dan Ilmu Komputer*, *2*(3), 69–76.

[5] Firdlous, D. A., Andrian, R., & Widodo, S. (2022). Sentiment Analysis Public Twitter on 2024 Election using the Long Short Term Memory Model. *SISTEMASI : Jurnal Sistem Informasi*, *12*(1), 52–60.

[6] Tan, K. L., Lee, C. P., & Lim, K. M. (2023). A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research. *Applied Sciences (Switzerland)*, *13*.

[7] Zain, H. H., Awannga, R. maulana, & Rahayu, W. I. (2023). Perbandingan Model Svm, Knn Dan Naïve Bayes Untuk Analisis Sentiment Pada Data Twitter: Studi Kasus Calon Presiden 2024. *JIMPS : Jurnal Ilmiah Mahasiswa Pendidikan Sejarah*, *8*(3), 2083–2093.

[8] Suryani, Fayyad, M. F., Savra, D. T., Kurniawan, V., & Estanto, B. H. (2023). Sentiment Analysis of Towards Electric Cars using Naive Bayes Classifier and Support Vector Machine Algorithm. *Public Research Journal of Engineering, Data Technology and Computer Science*, *1*(1), 1–9.

[9] Kanakaraddi, S. G., Chikaraddi, A. K., Gull, K. C., & Hiremath, P. S. (2020). Comparison Study of Sentiment Analysis of *Tweet*s using Various Machine Learning Algorithms. *Proceedings of the 5th International Conference on Inventive Computation Technologies, ICICT 2020*, 287–292.

[10] Lidinillah, E. R., Rohana, T., & Juwita, A. R. (2023). Analisis Sentimen Twitter Terhadap Steam Menggunakan Algoritma Logistic Regression dan Support Vector Machine. *TEKNOSAINS : Jurnal Sains, Teknologi Dan Informatika*, *10*(2), 154–164.

[11] Sinanto Ate, I., & Nuraminah, A. (2022). Komparasi Algoritma *Feature Selection* Pada Analisis Sentimen Review Film. *Jurnal Ilmiah Teknik Informatika Dan Komunikasi*, 2(2), 96–102.

[12] Madhumathi, R., Kowshalya, A. M., & Shruthi, R. (2022). Assessment of Sentiment Analysis Using Information Gain Based Feature Selection Approach. *Computer Systems Science and Engineering*, 43(2), 849–860.

[13] Anggraeni, A., Cholissodin, I., & Marji, M. (2021). Sentimen Analisis Layanan Produk Indihome menggunakan Information Gain dan Metode K-Nearest Neighbor. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *5*(8), 3616–3624.

[14] Rusman, J., Haryati, B. Z., & Michael, A. (2023). Optimisasi Hiperparameter Tuning pada Metode Support Vector Machine untuk Klasifikasi Tingkat Kematangan Buah Kopi. *Jurnal Komputer Dan Informatika*, *11*(2), 195–202.